doi:10.1111/evo.14

The Camouflage Machine: Optimizing protective coloration using deep learning with genetic algorithms

John G. Fennell,[1,2] Laszlo Talas,[1] Roland J. Baddeley,[1] Innes C. Cuthill,[3] and Nicholas E. Scott-Samuel[1]

[1]School of Psychological Science, University of Bristol, Bristol, UK

[2]E-mail: john.fennell@bristol.ac.uk

[3]School of Biological Sciences, University of Bristol, Bristol, UK

Received March 13, 2020

Accepted December 23, 2020

Evolutionary biologists frequently wish to measure the fitness of alternative phenotypes using behavioral experiments. However, many phenotypes are complex. One example is coloration: camouflage aims to make detection harder, while conspicuous signals (e.g., for warning or mate attraction) require the opposite. Identifying the hardest and easiest to find patterns is essential for understanding the evolutionary forces that shape protective coloration, but the parameter space of potential patterns (colored visual textures) is vast, limiting previous empirical studies to a narrow range of phenotypes. Here, we demonstrate how deep learning combined with genetic algorithms can be used to augment behavioral experiments, identifying both the best camouflage and the most conspicuous signal(s) from an arbitrarily vast array of patterns. To show the generality of our approach, we do so for both trichromatic (e.g., human) and dichromatic (e.g., typical mammalian) visual systems, in two different habitats. The patterns identified were validated using human participants; those identified as the best for camouflage were significantly harder to find than a tried-and-tested military design, while those identified as most conspicuous were significantly easier to find than other patterns. More generally, our method, dubbed the "Camouflage Machine," will be a useful tool for identifying the optimal phenotype in high dimensional state spaces.

KEY WORDS: Camouflage, deep learning, genetic algorithms, optimization, protective coloration.

The study of coloration has illuminated many important phenomena in evolutionary biology such as speciation, hybridization, the rate and direction of selection, dominance, linkage, sexual selection, mimicry and, more generally, adaptation (Cuthill et al. 2017). However, color patterns (visual textures with multiple colors) are difficult to characterize. While a color can be represented in a relatively low-dimensional space based on spectral characteristics, photoreceptor sensitivities, or psychophysical measurements (Renoult et al. 2017), a pattern (a combination of visual texture and one or more colors) is a high-dimensional attribute (Osorio and Cuthill 2015; Stoddard and Osorio 2019). The problem of characterization is particularly acute when the interest is in a color pattern shaped by not only the habitat, but also the perception of signal receivers with different visual systems. This will be the case for both camouflage and signals. For example, a poison dart frog (*Dendrobates* spp.) may be predated by birds, reptiles or mammals, each of which have different types of color vision; furthermore the same color pattern can function as either warning coloration or camouflage, dependent upon viewing distance and the predator's visual acuity (Barnett et al. 2018). Quantifying even a single color pattern may require representation in multiple perceptual spaces, each appropriate for a different observer with a different visual system (Caro 2014; Renoult et al. 2017). Nevertheless, characterization is just the starting point for the even greater problem that the scientist faces: searching a high-dimensional space for an optimal solution that can be compared to that, or those, of evolution. Identifying the match, or mismatch, between the observed phenotypes and the optima predicted under different constraints is a key tool in the study of adaptation.

Here, we show how residual deep neural networks (RDNNs) (Abadi et al. 2016), combined with genetic algorithms (GAs), can be harnessed to classical psychophysical techniques to find different optima in high-dimensional spatiochromatic spaces. This allows us to determine the best color pattern for concealment, or for signaling, in a given habitat for a given observer. To illustrate the context and better understand the depth of the problem consider, for example, the study of animal camouflage. Typically, research has experimentally tested a small set of pattern types relevant to a specific functional hypothesis, or has identified ecological correlates of extant patterns, that is patterns seen in nature (Caro 2014; Stevens 2015; Ruxton et al. 2018; Cuthill 2019). For example, a comparative study of coat colors in felids shows a correlation with ecology (Allen et al. 2011), but not whether the observed patterns are the optima for the associated habitats or constrained by either the pattern-generation mechanisms or pigments available to mammals. Such studies necessarily omit possible patterns that evolution has not realized because of phylogenetic or developmental constraints, and so cannot identify the influence (if any) of such constraints. Furthermore, without comparison to the optimal pattern(s), it is hard to identify the extent to which an observed pattern is subject to trade-offs with other functions such as, for example, thermoregulation or UV-protection (Penacchio et al. 2015; Cuthill et al. 2017).

Defining a framework that could characterize patterns in terms of their visibility in a given context to a given viewer, in an efficient and biologically relevant way, would be an exceptionally useful research tool. As well as providing insight into the evolution of animal camouflage, it would also allow the assessment of whether the signals that animals use to display, variously, their qualities to mates or unprofitability to predators, are optimized for conspicuity. These may be subject to trade-offs that render maximal conspicuity suboptimal and/or favor tuning of the signal to particular receivers at particular distances (Bohlin et al. 2008; Barnett and Cuthill 2014; Barnett et al. 2017, 2018). In the human domain, our method may be useful in the development of bespoke camouflage for specific contexts, maximizing the visibility of warning signs, or helping to reduce visual clutter due to infrastructure.

The main purpose of this article is to propose and test a new method that can identify the best patterns for a given environment, to further our understanding into whether selection has been able to realize optimal solutions for animal coloration. Depending on context and requirements, the method is applicable for finding patterns that will be effective either for camouflage or to be highly conspicuous. Historically, methods used to evaluate patterns tend to be based on binary comparison (is the target in picture A or B?) or to measure detection speed and accuracy, typically on computer screens. This is useful if there are only a few patterns to compare, but if the aim is not to constrain the space of possible patterns artificially then this approach is inadequate. Our method proposes gathering data, provided by human participants, on a subset of the parameter space and then, using RDNNs (Abadi et al. 2016), to interpolate between pattern and detection time pairs to predict the detection time for empirically untested patterns.

To make the method highly applicable to real-world scenarios, we constructed naturalistic stimuli and, for realism, projected them on a screen large enough to fill the visual field. We used backgrounds taken from photographs of both temperate forest and scrub desert with foreground occlusion layers and targets inserted into the scenes using blue screening ("chroma key"), a method commonly employed in the film industry. We were also keen that the textures on the targets that we used had biological plausibility. To achieve this, we used two-component reaction-diffusion equations. These systems, originally proposed by Turing 1952 and Murray 2003, consist of semilinear parabolic partial differential equations capable of creating a vast array of textures including the camouflage patterns of animals (Allen et al. 2011, 2013). Textures were color mapped using one color (represented as an RGB triplet) for each of the two components, creating two-color, natural-looking patterns. We have tested our method using two color vision systems: trichromatic, representative of the human visual system (but which also includes catarrhine and some platyrrhine monkeys) and dichromatic, representing most other mammals, which are red-green color blind (Jacobs 1993). Our method could be used for other visual systems and, indeed, may be particularly useful here, where testing of subjects is technically more demanding and necessarily more time consuming. Most insects are trichromats, but with ultraviolet, "blue" and "green" photoreceptors; birds are tetrachromats, spanning insect and human spectral wavebands; and reptiles, amphibians, and fish show diverse types of tetra-, tri-, di-, and monochromacy (Kelber et al. 2003). As long as the stimuli can be displayed as desired (e.g., containing UV content) and the stimulus-space sampled adequately, our method can interpolate to estimate responses to unseen stimuli.

Experiments involved participants finding an object with a particular color pattern, from here on referred to as a target. Targets in our main experiments were constructed using nine dimensions (three for each of the two colors and three for texture), resulting in a parameter space containing a total of 6.18×10^{17} possible patterns. Since our parameter space was so large, we were unable to select targets exhaustively or randomly with sufficient diversity. Therefore, we implemented a GA to optimize the color and texture parameters, based on participants' responses trial by trial, for hardest or easiest to see stimuli (Mitchell 1998). Our first three experiments were pilot experiments conducted to validate the GA using an increasing number of optimized dimensions: the first experiment optimized for targets with single trichromatic

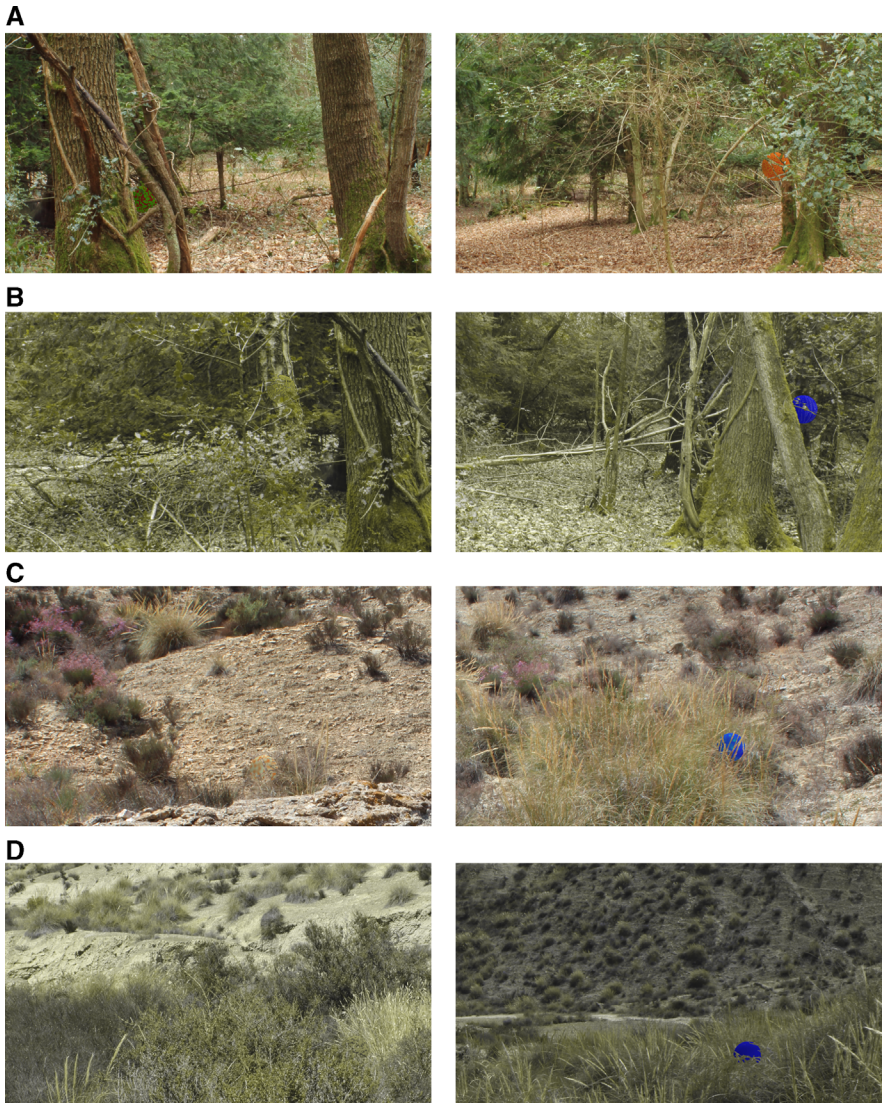

Figure 1. Find the spheres. Examples of experimental stimuli shown in experiments 4a-d. From top to bottom each row depicts the following conditions: trichromatic temperate forest; dichromatic temperate forest; trichromatic semiarid desert; dichromatic semiarid desert. Columns illustrate examples of hard (left) and easy to see (right) targets. For locations of the hard to see spheres in the left hand column see Figure S1 in Supporting information.

colors; the next experiment tested the optimizer with greyscale reaction-diffusion textures; and the final pilot optimized for two colors, but using a fixed pattern. Our hypothesis was that, over experimental generations of the GA, the reaction times to targets would gradually increase or decrease depending on whether targets were optimized for camouflage or conspicuity, respectively. Analysis using General Linear Mixed Models (GLMMs) showed support for a working GA and we then proceeded with our main experiment.

The main experiment followed a 2×2 design with two types of backgrounds (temperate forest or semiarid desert) and two color vision conditions (trichromatic or simulated dichromatic); examples of the stimuli are illustrated in Figure 1. The results of the main experiment were used to train RDNNs which we then used to predict reaction times (a measure of difficulty) for a far greater number of patterns than had been observed by the human participants. A final experiment was conducted to assess whether the method had produced an effective camouflage, by testing the patterns created against a tried-and-tested military pattern: Disruptive Pattern Material (DPM). DPM was a camouflage used by British Armed Forces for over 40 years and proven effective in temperate forest areas (Wynne 1972).

We call our method The Camouflage Machine, where "machine" is used to identify an effective method (or algorithm) for calculating a function that emphasizes the input/output relationship of natural images to optimized camouflage patterns, rather

than the particular choice of steps used in the process. Use of the term machine in this way dates back to at least Jevons (1870) and probably most famously to Turing (1937). The Camouflage Machine is a complete pipeline for generating biologically realistic color patterns, assessing their detectability against specified backgrounds for specified visual systems, predicting the detectability of vastly more, unseen, patterns (using deep learning) and evolving new and better patterns (using reaction-diffusion equations). This allows determination of the optimal coloration for a specified visual system, background, pattern-generation mechanism and function (concealment or signaling), thus, helping determine the constraints under which natural color patterns have evolved. Furthermore, the method could be generalized to tackle other sensory modalities (e.g., sound) where the stimulus space can be characterized but the range of possible stimuli greatly exceeds those which could be tested empirically.

Materials and Methods

PARTICIPANTS

A total of 95 participants (71 females, 24 males) were recruited from the University of Bristol. All participants had normal or corrected-to-normal vision. Informed consent was obtained from all participants as stated in the Declaration of Helsinki. All experiments were approved by the Ethics Committee of the University of Bristol's Faculty of Science.

STIMULI

The creation of stimuli used the same approach as Fennell et al. (2019). Stimuli were created from three layers. A background layer consisted of a natural scene taken from one of two locations: Leigh Woods (North Somerset, UK, 2°38.6' W, 51°27.8' N) and Tabernas Desert (Almería, Spain, 2°41.3' E, 37°02.9 N). A foreground layer was created by using a large blue cotton screen (1.8 × 2.8 m) shifted across the same background. All natural images were captured with a Nikon D90 digital SLR camera (Nikon Corp., Tokyo, Japan) at a 4288 × 2848 pixel resolution, mounted on a tripod. The natural images captured for both background and foreground were cropped to 1920 × 1080 pixels prior to further processing. Between the foreground and background, a target layer was constructed from colors and textures (see below). We preprocessed the blue screen images to create a mask for all possible locations for the centers of targets. The derived mask allowed rapid location selection and the introduction of occlusion in the foreground. A bespoke program, written using the Psychtoolbox-3 extension (Brainard 1997) for Matlab (Mathworks 2015), was used to construct and present the stimuli, and to collect experimental data.

During all experiments, stimuli were dynamically constructed from the three layers. Backgrounds were randomly chosen from a pool of 64 images (per geographical location). Using the associated mask, a location for the target was randomly selected. Based on the number of backgrounds and potential target positions there were a very large number of potential unique stimuli.

The target was always a sphere with a radius of 64 pixels. After applying colors and texture (specific to the experiments described below), we added pseudorealistic shading to produce a spherical look. The shape of a sphere was chosen as it was straightforward to create and provide with a scene-appropriate shading. Maintaining the spherical shape throughout the experiments managed the potential problem of a target appearing different from varying angles.

Where dichromatic images were used, representations of the stimuli were created using the protan equation (Viénot et al. 1999), which simulates a trichromatic representation of an image perceived by a protanopic dichromat.

TEXTURES

To generate biologically plausible textures, we implemented the Gray–Scott model of reaction diffusion (Pearson 1993). Full details are provided in Supporting information.

OPTIMISATION

To optimize the color and texture parameters, we used a GA based on participants' responses. Parameters for the first generation of stimuli were randomly selected from the parameter spaces for each of the experiments identified below (e.g., three for each color and three for each pattern), and the time taken to identify the stimulus recorded (fitness). A new generation was generated every 50 trials, where individual samples were selected with a GA using tournament-based selection and tournament size of 4. Tournament-based selection is an efficient method of selecting an individual from a population of individuals in a GA (Goldberg and Deb 1991; Blickle and Thiele 1996; Mitchell 1998). Tournament-based selection involves running "competitions" between members of a population, chosen at random, where the winner of each competition, the member with the best fitness, is selected for crossover. A larger tournament size reduces the probability that weak individuals will be selected (since there is a higher probability that a stronger individual is also in that tournament), thereby increasing selection pressure. Offspring, through the crossover process, received 50% of genes from each parent, for example, the best two individuals from the tournament, selected randomly. This was followed by a mutation rate of 10%, which assigned random values (mutations) to genes, randomly. The GA was run for various numbers of generations dependent upon the experiments described below.

GENERAL PROCEDURE

Images were projected on to a 1900 × 1070 mm screen (Euroscreen, Halmstad, Sweden) from 3100 mm using a 1920 × 1080 pixel HD (contrast ratio 300,000:1) LCD projector (PT-AE7000U; Panasonic Corp., Kadoma, Japan). For Yxy measurements of projected colors, see Table S1 in Supporting information. Participants sat 2 m away from the display screen, so that the experimental stimulus subtended a visual angle of 50.89° by 28.59° and the target sphere 3.64°. A central fixation cross on a mid-grey background was displayed for 2 s prior to stimulus onset. Participants were asked to indicate on which side of the screen they saw the target, using the left and right shift keys on a keyboard. Each trial had a 10 s timeout; if this was reached, the experiment automatically advanced. The intertrial interval was set to 2 s. Failure to respond was recorded as a failure and the experiment moved on the next stimulus. Reaction times were recorded to the nearest millisecond and errors indicating choice of the wrong side of the screen were logged.

EXPERIMENTS

For each experiment (unless stated otherwise), half of the participants saw targets optimized for increasing difficulty, while the other half were presented with targets optimized for increased visibility. Occlusion levels were maintained between 25 and 50% of the target, chosen randomly from a uniform distribution. Experiment 1 had 10 participants (eight females, two males) with targets of a single color presented on temperate forest backgrounds in trichromatic color, optimized over 500 trials. Experiment 2 had 10 participants (eight females, two males) featuring monochrome stimuli with evolving textures presented on temperate forest backgrounds, optimized over 500 trials. Experiment 3 had 10 participants (seven females, three males) who were shown targets with a fixed disruptive texture and two colors against a temperate forest background in trichromatic color, optimized over 1000 trials. In this experiment, all participants were shown targets optimized to be hard to see.

After we confirmed that the optimizer worked, the main experiment (Experiment 4) followed a 2 × 2 design with two types of backgrounds (temperate forest or desert scrub) and two color vision conditions (trichromatic or dichromatic). Forty participants (seven males, 33 females) were randomly divided between the four conditions. Each participant completed 1000 trials.

DEEP NEURAL NETWORKS

While the stimuli were generated and the experiments run using Matlab programs, the RDNNs were written in Python 3 (Python Software Foundation, Wilmington, DE) using neural network API Keras (Chollet et al. 2015). Each network was of the same configuration and consisted of an input layer, a number of residual blocks, and an output layer. The input layer was of 22 units,

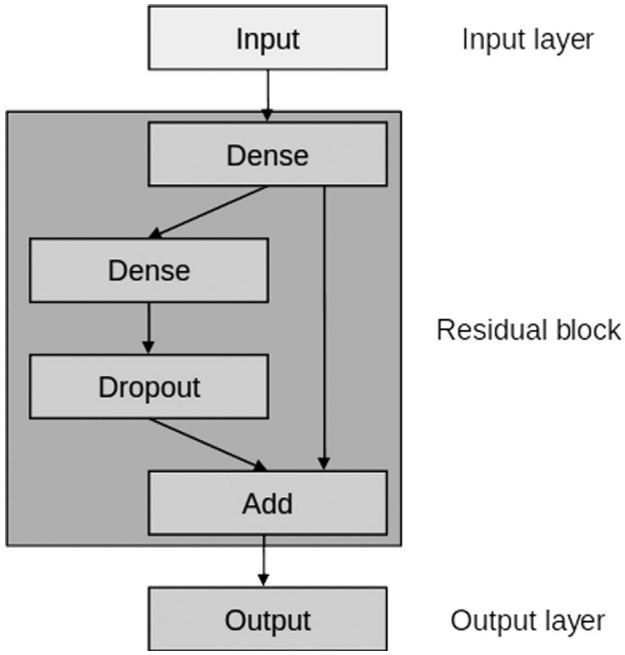

Figure 2. Schematic illustration of the residual deep neural network used in the study.

comprising three dimensions for the pattern color representing substance A, as described above for the Gray–Scott model; three dimensions for the pattern color representing substance B in the Gray–Scott model; three dimensions for the texture; a dimension for level of occlusion; a two element one-hot array to indicate the optimization (hardest or easiest); and a 10 element one-hot array to identify the participant. A one-hot array is a 1 × N array used to distinguish each category in a set (size N) from every other category in the set. The vector consists of zeros in all vector locations except for a single 1 in the location used to uniquely identify the category. Input colors, both trichromat and simulated dichromat, were represented as RGB triplets, with simulated dichromat values consisting of R and G channels of the same value. An alternative color space, such as CIELab or HSV, could have been used, but as neural networks form their own internal representations of distances (Rafegas and Vanrell 2018) the choice of color space is irrelevant.

Residual blocks, each comprised two dense layers, a dropout layer and a summation layer, containing 768 units each, and an output layer consisting of a single variable representing difficulty as reaction time (Fig. 2). We used the built in "rmsprop" optimizer from Keras with the "mean squared error" loss function, on difficulty, to train the networks, based on a batch size of 128 for 500 epochs.

To establish the number of residual blocks to use, networks were trained with one, two, four, and six residual blocks. When training a network model, a proportion of the dataset is "held-out" for validation. The training loss is the error on the training

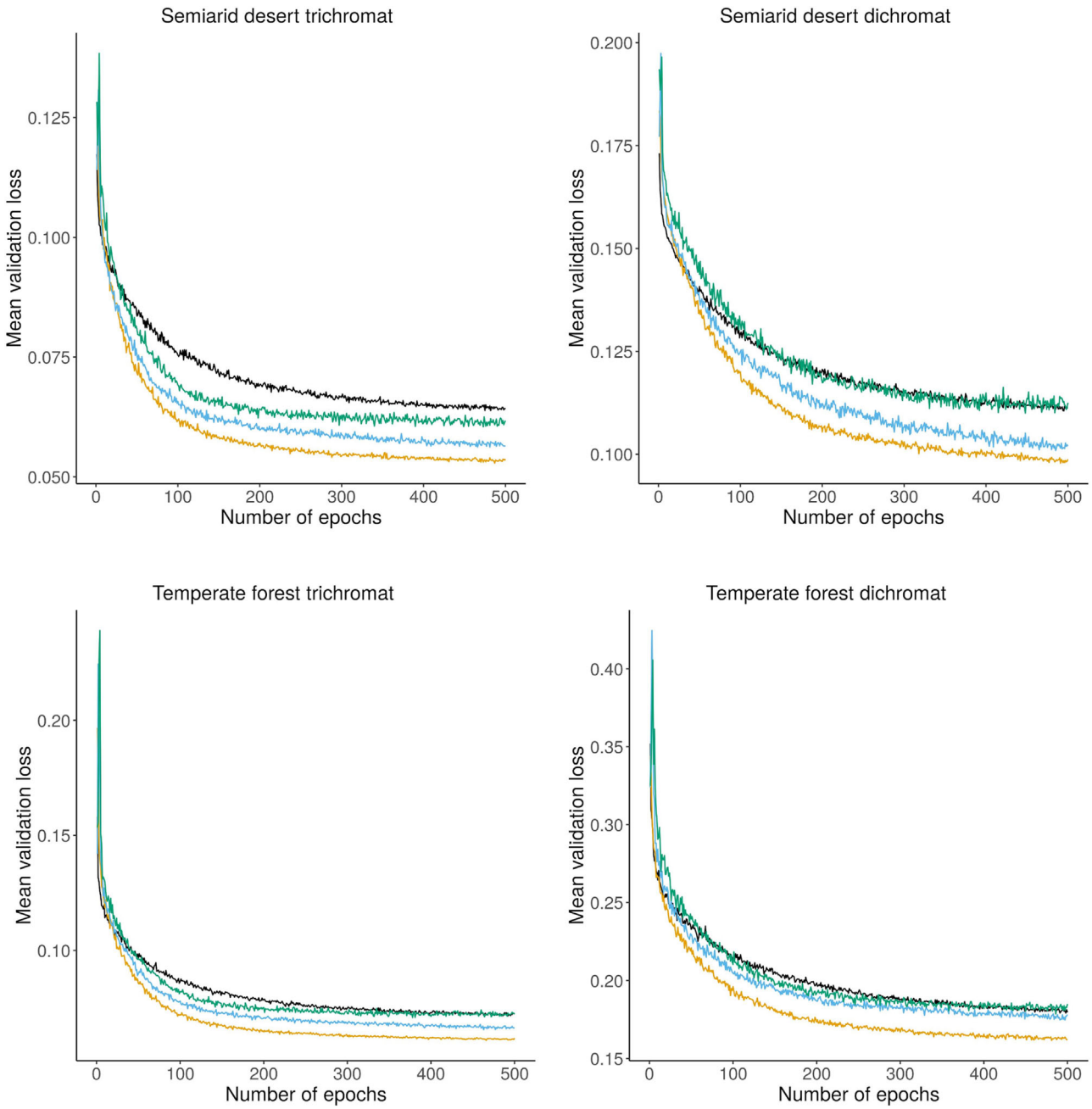

Figure 3. **Mean validation losses between neural networks with one, two, four, or six residual blocks across 500 training epochs were compared to identify the network with the lowest loss, for each of the four experimental conditions. Networks with the best fit, that is, lowest losses (per condition) were used to generate patterns for camouflage and conspicuity.**

set of data, in the present case calculated using mean squared error, while the validation loss is the error, calculated in the same way, after running the held-out validation set through the trained network. As the number of epochs increases, it is expected that both the validation and training error will drop. Put simply, if validation losses are compared across different models trained with the same data, the model with the lower loss would be preferred. Here, mean validation losses were calculated for 100 bootstrapped neural networks after 500 training epochs using mean squared error (Fig. 3). Statistics to compare the effects of residual block number were calculated using random permutation tests, based on 100,000 resamples. *P*-values were adjusted for multiple comparisons with False Discovery Rate (Benjamini and Hochberg 1995; Bates et al. 2015). We found that neural networks with two residual blocks produced significantly lower error rates compared to networks with one or six residual blocks, in all four experimental conditions (Table 1). While networks with two residual blocks produced significantly lower error rates

Table 1. Comparisons of mean validation losses for neural networks with two versus one, four and six residual blocks in all four experimental conditions.

Condition	Comparison	*P* value
Temperate forest trichromat	2 vs. 1	<0.0001
	2 vs. 4	0.0054
	2 vs. 6	<0.0001
Temperate forest dichromat	2 vs. 1	<0.0001
	2 vs. 4	<0.0001
	2 vs. 6	<0.0001
Semi-arid desert trichromat	2 vs. 1	<0.0001
	2 vs. 4	0.0518
	2 vs. 6	<0.0001
Semi-arid desert dichromat	2 vs. 1	<0.0001
	2 vs. 4	0.1694
	2 vs. 6	<0.0001

compared to networks with four residual blocks in temperate forest conditions, the difference was not significant in semiarid desert conditions. Therefore, applying Occam's razor, we used networks with two residual blocks as it was simpler.

VALIDATION EXPERIMENT

The top 25 hardest and easiest to find patterns predicted by our method from the temperate forest trichromat condition were paired with 25 DPM and 25 averaged patterns (Fig. S2 in Supporting information) for an experimental run with human participants. One run contained each pattern four times in a random order (totaling 100 trials), supplemented by four randomly selected patterns from each condition presented at the start as practice trials. We recruited 25 participants (15 females, 10 males) for the validation experiment, where each run was presented to a single participant. In all other aspects, the experiment was identical to those described above.

Results

The three pilot experiments confirmed that the GA was capable of optimizing target color and texture for both concealment and high visibility. GLMMs showed that trials became significantly harder over time when optimizing for concealment, while optimizing for visibility yielded easier to find targets (Table 1). The effects of trial number on log-transformed reaction times were analyzed by fitting general linear mixed models using the lme4 package (Bates et al. 2015) in R (R Core Team 2015). Nested models were compared using the change in deviance on removal of the fixed variable for GA generations. A positive estimate coupled with a significant *P*-value suggested that targets became harder to see over the course of the experiment, while negative estimates indicated that targets became easier to see. It should be noted that estimates and standard deviations are presented as log-transformed reaction times. For example, an estimate of 1.86e-4 indicates that the target in the final trial was approximately one second harder to find than the target in the initial trial. In the main experiment, the optimizer produced significantly harder/easier results according to settings (see Table 2, experiments 4A-D), except in the dichromat desert condition optimized for easiest to see targets ($P = 0.5321$); we address this in the discussion below.

RDNNs were implemented in Keras 2.1.2 (Chollet et al. 2015) utilizing the neural network library TensorFlow 1.5.0 (Abadi et al. 2016) and were trained with all of the samples collected from the main experiment. To provide for a measure of precision in our predictions (an estimate of standard error of the mean), we created 100 bootstraps of our networks for each of the four conditions. The bootstrap method is a test or metric that uses random sampling with replacement. The bootstrap method allows assignment of measures for precision, defined here in terms of standard error of the mean and is particularly useful when the value of interest is, as in the present case, a complicated function (Efron and Tibshirani 1994). By averaging the bootstrapped networks' predictions we calculate both a data-dependent smoothing of the reaction time function and an estimate of our certainty of its estimate. Each network was trained on a random sample of 90% of the data and validated with the remaining 10%.

Predicting the full parameter space poses a computational challenge due its vastness. We therefore created 100 "artificial observers" based on each of the 100 models, using a similar GA to the one discussed above. The artificial observers were used to generate 1000 optimized samples each. Averaged reaction times for the combined 100,000 samples were obtained using all 100 models. The top 25 patterns that were identified for each condition and optimization setting are illustrated in Figure S4 in Supporting information. Figure 4 shows the mean predicted reaction times of the top 25 patterns per condition by artificial observers.

The top 25 (both optimized for hardest and easiest) patterns identified in the trichromatic temperate forest condition were tested together with two additional control patterns: British DPM camouflage (Wynne 1972) and the mean color obtained by averaging across all woodland backgrounds. DPM was used by the British Armed Forces, and many other nations, for over 40 years. We therefore considered it an appropriate control that avoids political sensitivities created by comparisons to any current military patterns. Furthermore, the average color of the background was khaki, which has been used by numerous militaries (including the British) from the late 19th century, making it also an important control pattern. Statistics were obtained using GLMMs, where the model, with the effect of treatment included, provided a significantly better fit than one without it (Δdeviance = 65.848, d.f. = 3, $P = 3.304\text{e-}14$). Post-hoc analysis (Tukey HSD)

Table 2. Details of the General Linear Mixed Model analysis to determine whether the Genetic Algorithm is effective, for all experiments. Effectiveness was established by comparing Δdeviance between models with and without the fixed variable for GA generations.

Experiment	Color	Optimized for	N	ΔDeviance	Df	*P*	Estimate	Std
Experiment 1	Trichromat	Hardest	5	17.293	1	3.20×10^{-05}	1.86×10^{-04}	4.46×10^{-05}
		Easiest	5	22.510	1	2.09×10^{-06}	-1.38×10^{-04}	2.89×10^{-05}
Experiment 2	Monochrome	Hardest	5	11.552	1	6.77×10^{-04}	3.17×10^{-04}	9.32×10^{-05}
		Easiest	5	317.050	1	$<2.20 \times 10^{-16}$	-1.14×10^{-03}	6.21×10^{-05}
Experiment 3	Trichromat	Hardest	10	101.520	1	$<2.2 \times 10^{-16}$	1.47×10^{-04}	1.46×10^{-05}
Experiment 4a	Trichromat	Hardest	5	10.771	1	0.001031	6.53×10^{-05}	1.99×10^{-05}
		Easiest	5	16.633	1	4.54×10^{-05}	-5.34×10^{-05}	1.31×10^{-05}
Experiment 4b	Dichromat	Hardest	5	47.902	1	4.48×10^{-12}	2.10×10^{-04}	3.03×10^{-05}
		Easiest	5	16.612	1	4.59×10^{-05}	-6.34×10^{-05}	1.56×10^{-05}
Experiment 4c	Trichromat	Hardest	5	7.565	1	0.005951	4.72×10^{-05}	1.71×10^{-05}
		Easiest	5	156.360	1	$<2.2 \times 10^{-16}$	-1.59×10^{-04}	1.26×10^{-05}
Experiment 4d	Dichromat	Hardest	5	5.160	1	0.02317	4.83×10^{-05}	2.13×10^{-05}
		Easiest	5	0.390	1	0.5321	-9.02×10^{-06}	1.44×10^{-05}

Positive estimates indicate an increase in reaction time, that is, patterns became significantly harder to see (optimized for camouflage), while negative estimates show a decrease in reaction time, that is, easier to see (optimized for conspicuity). Experiments 1–3 are pilot experiments to test optimization for single trichromatic colors, greyscale reaction-diffusion textures, and a single, fixed pattern with two colors, respectively. Experiments 4a–d are the main experiment where both colors and textures are optimized.

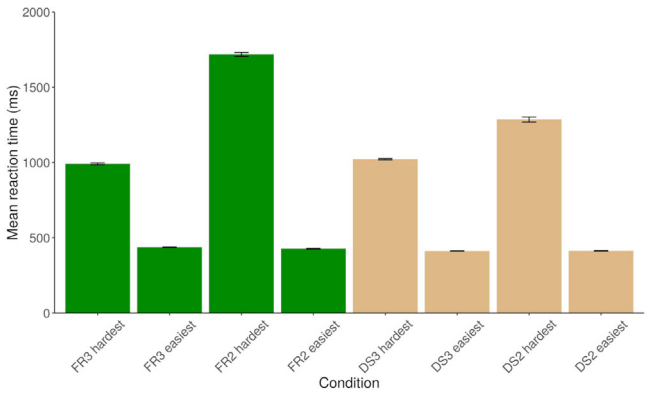

Figure 4. Mean predicted reaction times of the top 25 patterns per condition identified by the Camouflage Machine. Error bars are standard error of the mean. FR3: trichromatic temperate forest; FR2: dichromatic temperate forest; DS3: trichromatic semiarid desert; DS2: dichromatic semiarid desert.

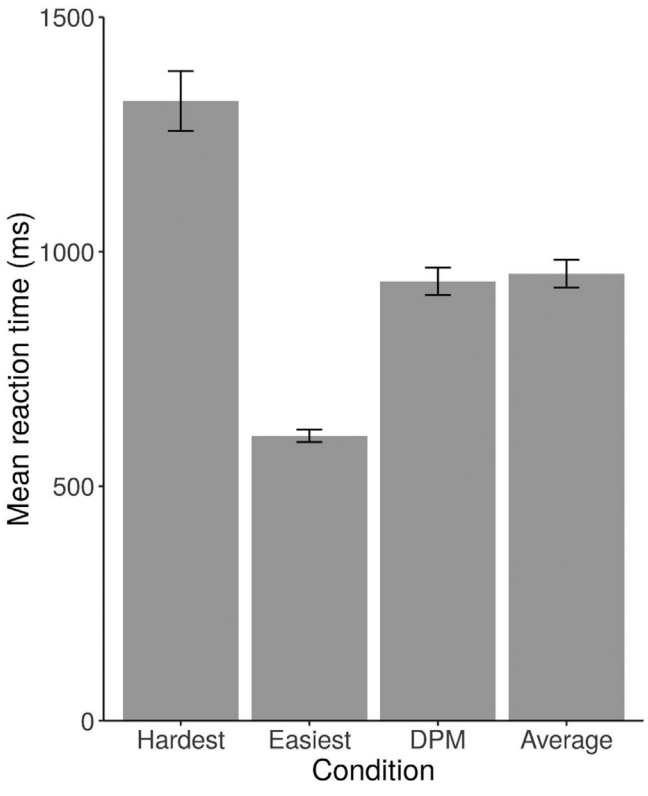

Figure 5. Mean reaction times to conditions tested in the validation experiment. Error bars are standard error of the mean.

showed clearly (see Fig. 5) that the hardest patterns identified by our method were significantly harder to detect than DPM ($P = 0.0256$) and the average color ($P = 0.0474$). Similarly, the easiest patterns according to our method were significantly easier to detect than DPM ($P < 0.001$) and the average color ($P < 0.001$).

Discussion

Evolutionary biologists frequently aim to measure the fitness, or some surrogate currency, of different phenotypes using behavioral measures, for example, survival, foraging success, detection, and attractiveness. However, for complex phenotypes, such as coloration, the state-space can be vast. There are two common solutions: one is to limit the experiment to a small number of

observed phenotypes, for example, melanic versus nonmelanic (Karpestam et al. 2014; Walton and Stevens 2018), or discrete variants in mimetic accuracy (Bain et al. 2007; Kikuchi and Pfennig 2010). The other solution is to abstract the problem to a range of simple stimuli that capture the essence of the question but do not attempt to mimic reality, for example, experiments with artificial prey to investigate the evolution of aposematism (Lindström et al. 1999), mimicry (Kazemi et al. 2014), or camouflage (Barnett and Cuthill 2014). Although the use of computer displays, either with human participants or nonhuman animals in an operant (reinforcement-based) paradigm, can reduce the time costs and so expand the range of phenotypes evaluated, it is still of the order of 100's not the millions that would ideally be investigated. The method presented here, named the Camouflage Machine, augments participant responses with AI, to vastly increase the scope of any such investigation.

The Camouflage Machine successfully identified patterns that were better, in terms of camouflage, than an existing military camouflage pattern and the average background color, commonly regarded as a good solution for concealment (Fennell et al. 2019). The Camouflage Machine provides an effective and efficient way to search very large parameter spaces to establish optimal patterns for camouflage, as well as conspicuity, in various environments. As illustrated by our simulated dichromat experiments and use of two very different backgrounds, the method generalizes to different color vision systems and across dissimilar environments. It is important to note that the Camouflage Machine need not identify a single best concealed/visible pattern, but can reveal multiple, similarly effective solutions. It is equipped to deal with the possibility that natural backgrounds contain sufficient heterogeneity that any method, including evolution, may not find a unique, best solution. Or that other factors (than camouflage) may determine the optimum within a range of similarly concealing solutions, for example the cost of pigment synthesis or trade-offs with thermoregulation. Supporting information Figure S4 illustrates that there is considerable visual variability between patterns within conditions, but not in terms of predicted difficulty (see Fig. 4).

In all cases, we found that the standard error of the predictions was less than 17 ms, constituting what we believe to be an indistinguishable perceptual difference in the context of visually complex and nonaffective stimuli (Paul et al. 2012; Ionescu 2016). It is also interesting to note that the predicted mean reaction times for the easiest to find patterns in each condition are equivalent. We believe this should be expected because a sufficiently salient stimulus in a complex scene should exhibit a pop-out effect (Treisman and Gelade 1980; McElree and Carrasco 1999; Henderson 2007). Although dichromat targets optimized for concealment were significantly harder to detect than trichromat ones, consistent with our previous findings on uniformly colored stimuli (Fennell et al. 2019), it should be stressed that our results are for trichromats using the visual information available to a dichromat, not natural dichromats neurophysiologically adapted to, and familiar with, using that level of information.

Previous studies have used evolving prey (Bond and Kamil 2002, 2006; Sherratt et al. 2007); however, an important benefit of the Camouflage Machine is that far larger parameter spaces can be explored, effectively predicting data for unseen stimuli. Although deep neural networks are capable of modelling a large parameter space, establishing optima in a principled way remains a challenge. While it is technically possible to exhaustively predict every possible pattern in a given parameter space, it is certainly impractical in a reasonable timescale for the space described in this study. Our solution involves combining GAs with deep neural networks, effectively training "artificial observers." Artificial observers allow us to be able to navigate the parameter space in a principled way and establish the hardest and easiest color pattern combinations within reasonable timescales. For example, the predicted two-color stimuli (optimized for concealment) were able to outperform an existing military pattern (Wynne 1972) developed specifically for the (temperate forest) environment used in the experiment. We found that our genetic optimizer worked well in producing increasingly harder or easier to find patterns However, in a single condition, dichromat stimuli optimized for conspicuity in the semiarid desert, an improvement in pattern detectability across all trials was not found. We believe the explanation for this stems from the narrower range of patterns that provide significant levels of concealment; in other words, the optimizer has to deal with a space where most patterns are highly visible and so was already at ceiling performance for the majority of trials.

The Camouflage Machine offers a novel and useful tool for scientific and societal applications. Biologists will be interested in testing various hypotheses about the coloration of animals in specific environments. For example, finding an optimal concealing pattern in an environment and comparing it to the camouflage of animals inhabiting that environment could reveal more about their visual ecology (Caro 2014; Cuthill et al. 2017).

The Camouflage Machine is also capable of contributing to the development of dual-purpose applications, where both concealment and visibility is simultaneously required. For example, distance-dependent defensive coloration (Bohlin et al. 2008; Barnett and Cuthill 2014), or providing different information to different viewers. Introducing viewing distance as a variable in the models would allow identification of patterns that are conspicuous close up, but become concealed at a distance (Barnett et al. 2017, 2018). While we deliberately limited ourselves to two colors and a simple (spherical) shape, it is clearly possible

to include a larger number of colors and more complex shapes. Added to this, measures other than reaction time can be used, for example, aesthetic preference.

Conclusions

The impracticality of using large arrays of patterns has previously been a limiting factor in camouflage research and studies of the adaptive value of coloration more generally (Cuthill et al. 2017). With the aid of GAs and deep neural networks, we have also demonstrated a novel approach to psychophysics, carried out using multiple dimensions. We have achieved this using a modest number of optimized samples collected from relatively few participants. Using the Camouflage Machine, it is possible to identify clusters of global optima efficiently for both concealment and conspicuity. The approach should generalize to other problems where evolutionary biologists want to measure the fitness, or some surrogate currency, of phenotypes using behavioral measures (e.g., survival, foraging success, detection, attractiveness), but where the state space of possible phenotypes is too large to evaluate directly.

AUTHOR CONTRIBUTIONS

RJB, ICC, and NESS conceived the project; JGF and LT designed and developed the method; JGF and LT collected the data; JGF and LT created the models and analysed the data; JGF led the writing with help from LT, ICC, RJB, and NESS; and all coauthors assisted with edits and approve publication.

ACKNOWLEDGMENTS

We thank two anonymous reviewers for comments on the manuscript; also Erik Stuchly, Siyan Ye, Khishika Naidoo, Frankie King, and Marija Miloradov, for their help during data collection and Jasmina Stevanov for her help with image acquisition in Spain. JGF and LT were supported by an EPSRC grant (EP/M006905/1) awarded to NES-S, RJB, and ICC.

DATA ARCHIVING

Archival location upon acceptance or statement that there is no data to be archived. The dyrad doi is https://doi.org/10.5061/dryad.31zcrjdjv

CONFLICT OF INTEREST

The authors declare no conflict of interest.

LITERATURE CITED

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. 2016. TensorFlow: large-scale machine learning on heterogeneous distributed systems. Cornell University. arXiv:1603.04467 [cs].

Allen, W. L., R. Baddeley, N. E. Scott-Samuel, and I. C. Cuthill. 2013. The evolution and function of pattern diversity in snakes. Behav. Ecol. 24:1237–1250.

Allen, W. L., I. C. Cuthill, N. E. Scott-Samuel, and R. Baddeley. 2011. Why the leopard got its spots: relating pattern development to ecology in felids. Proc. Royal Soc. B: Biol. Sci. 278:1373–1380.

Bain, R. S., A. Rashed, V. J. Cowper, F. S. Gilbert, and T. N. Sherratt. 2007. The key mimetic features of hoverflies through avian eyes. Proc. Royal Soc. B: Biol. Sci. 274:1949–1954.

Barnett, J. B., and I. C. Cuthill. 2014. Distance-dependent defensive coloration. Curr. Biol. 24:R1157–R1158.

Barnett, J. B., I. C. Cuthill, and N. E. Scott-Samuel. 2017. Distance-dependent pattern blending can camouflage salient aposematic signals. Proc. Royal Soc. B: Biol. Sci. 284:20170128.

Barnett, J. B., C. Michalis, N. E. Scott-Samuel, and I. C. Cuthill. 2018. Distance-dependent defensive coloration in the poison frog *Dendrobates tinctorius*, Dendrobatidae. PNAS 115:6416–6421.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting linear mixed-effects models using lme4. J. Stats. Software 67:1–48.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Royal Stat. Soc.: Series B 57:289–300.

Blickle, T., and L. Thiele. 1996. A comparison of selection schemes used in evolutionary algorithms. Evol. Computat. 4:361–394.

Bohlin, T., B. S. Tullberg, and S. Merilaita. 2008. The effect of signal appearance and distance on detection risk in an aposematic butterfly larva (*Parnassius apollo*). Anim. Behav. 76:577–584.

Bond, A. B., and A. C. Kamil. 2002. Visual predators select for crypticity and polymorphism in virtual prey. Nature 415:609–613.

———. 2006. Spatial heterogeneity, predator cognition, and the evolution of color polymorphism in virtual prey. PNAS 103:3214–3219.

Brainard, D. H. 1997. The psychophysics toolbox. Spat Vis 10:433–436.

Caro, T. 2014. Concealing coloration in animals by Judy Diamond and Alan B. Bond. Quart. Rev. Biol. 89:63–63.

Chollet, F. et al. 2015. Keras. GitHub. https://github.com/fchollet/keras

Cuthill, I. C. 2019. Camouflage. J. Zool. 308:75–92.

Cuthill, I. C., W. L. Allen, K. Arbuckle, B. Caspers, G. Chaplin, M. E. Hauber, G. E. Hill, N. G. Jablonski, C. D. Jiggins, A. Kelber, et al. 2017. The biology of color. Science 357:eaan0221.

Efron, B., and R. J. Tibshirani. 1994. An Introduction to the Bootstrap. CRC Press, Cleveland, OH.

Fennell, J. G., L. Talas, R. J. Baddeley, I. C. Cuthill, and N. E. Scott-Samuel. 2019. Optimizing colour for camouflage and visibility using deep learning: the effects of the environment and the observer's visual system. J. Royal Soc. Interface 16:20190183.

Goldberg, D. E., and K. Deb. 1991. A comparative analysis of selection schemes used in genetic algorithms. Pp. 69–93 *in* G. J. E. Rawlins, ed. Foundations of Genetic Algorithms, Elsevier, Amsterdam, The Netherlands.

Henderson, J. M. 2007. Regarding scenes. Curr. Dir. Psychol. Sci. 16:219–222.

Ionescu, M. R. 2016. Subliminal perception of complex visual stimuli. Rom. J. Ophthalmol. 60:226–230.

Jacobs, G. H. 1993. The distribution and nature of colour vision among the mammals. Biol. Rev. 68:413–471.

Jevons, W. S. 1870. On the mechanical performance of logical inference. Philosoph. Transact. Royal Soc. London 160:497–518.

Karpestam, E., S. Merilaita, and A. Forsman. 2014. Natural levels of colour polymorphism reduce performance of visual predators searching for camouflaged prey. Biol. J. Linn. Soc. 112:546–555.

Kazemi, B., G. Gamberale-Stille, B. S. Tullberg, and O. Leimar. 2014. Stimulus salience as an explanation for imperfect mimicry. Curr. Biol. 24:965–969.

Kelber, A., M. Vorobyev, and D. Osorio. 2003. Animal colour vision—behavioural tests and physiological concepts. Biol. Rev. 78:81–118.

Kikuchi, D. W., and D. W. Pfennig. 2010. Predator cognition permits imperfect coral snake mimicry. Am. Natural. 176:830–834.

Lindström, L., R. V. Alatalo, J. Mappes, M. Riipi, and L. Vertainen. 1999. Can aposematic signals evolve by gradual change? Nature 397:249–251.

Mathworks. 2015. Matlab 2015b. The MathWorks, Inc., Natick, MA.

McElree, B., and M. Carrasco. 1999. The temporal dynamics of visual search: evidence for parallel processing in feature and conjunction searches. J. Exp. Psychol. Hum. Percept. Perform. 25:1517–1539.

Mitchell, M. 1998. An Introduction to genetic algorithms. Reprint Edition. MIT Press, Cambridge, MA.

Murray, J. D. 2003. Mathematical biology II: spatial models and biomedical applications. 3rd ed. Springer-Verlag, New York, NY.

Osorio, D., and I. C. Cuthill. 2015. Camouflage and perceptual organization in the animal kingdom. Pp. 843–862 *in* J. Wagemans (Ed.), The Oxford handbook of perceptual organization. Oxford Univ. Press, Oxford, U.K.

Paul, E. S., S. A. J. Pope, J. G. Fennell, and M. T. Mendl. 2012. Social anxiety modulates subliminal affective priming. PLOS ONE 7:e37011.

Pearson, J. E. 1993. Complex patterns in a simple system. Science 261:189–192.

Penacchio, O., I. C. Cuthill, P. G. Lovell, G. D. Ruxton, and J. M. Harris. 2015. Orientation to the sun by animals and its interaction with crypsis. Functional Ecol. 29:1165–1177.

R Core Team. 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rafegas, I., and M. Vanrell. 2018. Color encoding in biologically-inspired convolutional neural networks. Vision Res. 151:7–17.

Renoult, J. P., A. Kelber, and H. M. Schaefer. 2017. Colour spaces in ecology and evolutionary biology. Biol. Reviews 92:292–315.

Ruxton, G. D., W. L. Allen, T. N. Sherratt, and M. P. Speed. 2018. Avoiding attack: the evolutionary ecology of crypsis, aposematism, and mimicry. 2nd ed. Oxford University Press, Oxford, U.K.

Sherratt, T. N., D. Pollitt, and D. M. Wilkinson. 2007. The evolution of crypsis in replicating populations of web-based prey. Oikos 116:449–460.

Stevens, M. 2015. Anti-predator coloration and behaviour: a longstanding topic with many outstanding questions. Curr. Zool. 61:702–707.

Stoddard, M. C., and D. Osorio. 2019. Animal coloration patterns: linking spatial vision to quantitative analysis. Am. Natural. 193:164–186.

Treisman, A. M., and G. Gelade. 1980. A feature-integration theory of attention. Cognit. Psychol. 12:97–136.

Turing, A. M. 1937. On computable numbers, with an application to the Entscheidungsproblem. Proc. London Mathematic. Soc. s2-42:230–265.

———. 1952. The chemical basis of morphogenesis. Philos. Transact. Royal Soc. London. Series B, Biol. Sci. 237:37–72.

Viénot, F., H. Brettel, and J. D. Mollon. 1999. Digital video colourmaps for checking the legibility of displays by dichromats. Color Res. Appl. 24:243–252.

Walton, O. C., and M. Stevens. 2018. Avian vision models and field experiments determine the survival value of peppered moth camouflage. Commun. Biol. 1:1–7.

Wynne, R. W. 1972. Assessment of effect of disruptively-patterned combat clothing on concealment. Royal Aircraft Establishment, Farnborough, UK.

Associate Editor: T. Chapman
Handling Editor: T. Chapman

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Examples of the hard to see experimental stimuli from the left hand column of Figure 1.

Figure S2. Top: The generated Gray-Scott space.

Table S1. Reference and measured values of projected colours using a Minolta CS-100A Luminance and Color Meter (Minolta Co., Ltd., Osaka, Japan).

Figure S3. Control stimuli used in the validation experiment: 25 British military (DPM) patterns (left) and average colour of temperature woodland backgrounds (right).

Figure S4. The top 25 hardest (left) and easiest (right) to see patterns identified by the Camouflage Machine for each condition tested.